Vision Transformer based System for Fruit Quality Evaluation

Tanushri Kumar (tanushri.skr@gmail.com)
Anna University Chennai College of Engineering Guindy
Shivani R
Anna University Chennai College of Engineering Guindy

Research Article

Keywords: Computer vision, Image processing, Quality assessment, Vision transformer

Posted Date: April 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1526586/v1

Abstract

Purpose Fruit quality assessment is one of the most pressing issues in the farming industry. Agriculturists would benefit significantly from the capacity to recognize the freshness of fruits, as it will allow them to optimize the harvesting stage and avoid reaping either underdeveloped or overdeveloped natural products. Productivity has decreased due to a lack of low-cost technology and equitable access. Despite large-scale agricultural mechanization in some parts of the country, most agricultural operations are still carried out by hand with simple instruments. The goal of this research is to automate the task of evaluating fruit quality.

Methods Transformers were first presented in the field of natural language processing, and they offer dramatic performance improvements over existing models in NLP like as LSTMs and GRU. The Vision Transformer, or ViT, is an image classification model that uses a Transformer-like design over image patches.

Results Vision transformer exhibited far superior results compared to CNN models. The models were evaluated on various metrics and the vision transformer model generated the highest accuracy compared to convolutional neural network models such as InceptionNet and EfficientNet.

Conclusions Vision Transformers exhibited superior performance in fruit quality evaluation compared to traditional CNN model approaches. For future works, this model can be used to develop an efficient quality control system Automation of the quality prediction process can greatly reduce food and agricultural produce wastage.

Introduction

Agriculture is essential to humans because it assures a steady supply of food and ensures food security for the population. Fruits, in particular, are often purchased by every family and are high in nutrients; hence, a constant supply and production are necessary to meet the demand of the world's rising population. As a result, the whole agri-food industry chain is facing rising problems, necessitating the application of new creative technologies in order to increase production. Fruits have been a staple of human diets since prehistoric times. Because of their excellent nutritional value, they offer a significant nutritional contribution to human well-being. It is necessary to ensure the quality of fruits ingested in all locations. Many fruit supply firms continue to ship improper fruit for consumption due to a lack of precision in the fruit sorting procedure performed by their employees. Computational technologies have been used for fruit recognition and other computer vision tasks (Kaur et al., 2015).

Because manual grading is time demanding, automation of the grading process through the use of computerized systems is seen to be the solution (Blasco et al., 2003). For this classification challenge, deep learning methods were applied using transfer learning, a machine learning method in which a model developed for one task is reused as the starting point for a model on a different problem. It is a popular approach in deep learning, where pre-trained models are used as the starting point for computer vision

and natural language processing tasks (S. Chakraborty et al., 2021). This has proved effective in carrying knowledge learned from general-domain, large-scale datasets to specific domains, where the amount of data available is limited. To accomplish this, a fruit freshness grading system that can distinguish various varieties of fruits from photographs acquired by any digital camera or smartphone from diverse locations can be developed. This device will assist us in determining the quality of fruits as well as developing a robotic orchard harvesting system.

Dataset

The primary concept of machine learning is collecting image datasets for training and model building. For the prediction of the quality of fruits, the dataset "IEEEFRUITSDATA_train&test" was used. The Dataset consists of 12,050 images comprising fruits such as Apple, Banana, Guava, Orange, Lime, and pomegranate categorized based on their quality as Good or Bad.

Materials And Methods

CNN-based approach

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning method that can take an input image, assign relevance (learnable weights and biases) to various aspects/objects in the image, and distinguish between them. Convolution is a linear mathematical action between matrices. A convolutional layer, a non-linearity layer, a pooling layer, and a fully-connected layer are among the layers of CNN. In machine learning issues, CNN has performed satisfactorily. The dataset under study was used to train deep learning models (S. Fan et al., 2020), namely EfficientNet and InceptionNet (Patil, 2018).

EfficientNet is a convolutional neural network architecture and scaling approach that uses a compound coefficient to consistently scale all the depth, breadth, and resolution parameters. Unlike conventional practice that arbitrarily scales these factors, the EfficientNet scaling method uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients (A. Kumar et al., 2021). The model was implemented using the transfer learning approach. A peak accuracy of 91% was achieved.

Inception v3 is an image recognition model that has been shown to attain greater than 78.1% accuracy on the ImageNet dataset. To achieve better accuracy, the InceptionNet model was implemented. The model is the culmination of many ideas developed by multiple researchers over the years. It is based on the original paper: "Rethinking the Inception Architecture for Computer Vision" (Szegedy et al., 2016). Convolutions, average pooling, max pooling, concatenations, dropouts, and fully linked layers are among the symmetric and asymmetric building components in the model. Batch normalization is done to activation inputs and is utilized extensively throughout the model. Loss is computed using Softmax. Due to its high accuracy on image classification problems, InceptionNet was utilized to develop a model. A peak accuracy of 94% was achieved. Though it exhibited better results on the test data compared to the EfficientNet model approach, the models took a long time to train, and its predictions on test data were not satisfactory enough.

Vision transformer

The Vision Transformer (ViT) has emerged as a viable alternative to convolutional neural networks (CNNs), which are the current state-of-the-art in computer vision and are widely employed in image identification applications. In terms of computing efficiency and accuracy, ViT models exceed the present state-of-the-art (CNN) by almost a factor of four.

Transformers are already capable of paying attention to regions that are far apart right from the starting layers of the network which is a significant gain the transformers bring over CNNs which have a finite receptive field at the start. One other advantage of transformer models is that they are highly parallelizable.

The attention mechanism enhances the crucial parts of the input data and fades out the rest. Self-attention module replaces the convolutional layer so that now the model gets the ability to interact with pixels far away from its location. The self-attention mechanism is a type of attention mechanism which allows every element of a sequence to interact with others and find out to whom they should pay more attention. An attention mechanism like self-attention can effectively solve some of the limitations of the Convolutional Networks. This distinct behavior is due to the inclusion of some inductive biases in CNNs, which can be used by these networks to comprehend the particularities of the analyzed images more rapidly, even if they end up limiting them and making it more difficult to grasp global relations.

The Vision Transformers, on the other hand, are free of these biases, allowing them to capture a global and wider range of relationships at the cost of more time-consuming data training. Input visual distortions such as adversarial patches or permutations were also significantly more resistant to Vision Transformers (Park et al., 2022).

Architecture-ViT model

The ViT model is made up of many Transformer blocks that employ layers. As a self-attention method, the MultiHeadAttention layer is applied to the sequence of patches. The Transformer blocks generate a [batch size, num patches, projection dim] tensor, which is then processed by a SoftMax classifier head to generate the final class probabilities output.

The initial stage in the model is to split an input image into a series of image patches. By projecting a patch onto a vector of size projection_dim, the PatchEncoder layer will linearly transform it. The projected vector is also given a learnable position embedding. These image patches are then sent via a linear projection layer that may be trained. This layer serves as an embedding layer, producing fixed-size vectors. The sequence of image patches is then linearly added using position embeddings to ensure that the images preserve their positional information. It injects crucial information about the image patches' relative or absolute positions in the sequence.

The 0th class is a principal element to note in the position embedding module. BERT's class token inspired the concept of the 0th class. This class, like the others, is learnt, although it does not originate

from its picture. Instead, the model design has it hardcoded. If the transformer is provided with the positioning data, it will not know what order the photos are in. The transformer encoder receives this sequence of vector pictures.

A Multi-Head Attention layer and a Multi-Layer Perceptron (MLP) layer make up the Transformer encoder module. The Multi-Head Attention layer divides inputs into several heads, allowing each head to develop varying levels of self-attention. All the heads' outputs are then combined and sent into the Multi-Layer Perceptron. Normalization layers (Layer Norm) are applied before each block using transformers, and residual blocks are applied afterward. Finally, the transformer encoder receives an additional learnable classification module (the MLP Head), which determines the network's output classes.

Model Details

The proposed work fine-tunes the google/vit-base-patch16-224-in21k a Vision Transformer (ViT) pre-trained on ImageNet-21k (14 million pictures, 21,843 classes) at 224x224 resolution in this case. The model is provided with images in the form of a series of fixed-size patches (resolution 16x16) that are linearly embedded. In order to train the model, the images must be converted to pixel values. A transformer's Feature Extractor accomplishes this by augmenting and converting the photos into a 3D Array that can be fed into the model (Dosovitskiyet al., 2020).

Data augmentation techniques (John B et al., 2002) were performed on images of the training set to improve the generalization ability of the model with the help of PyTorch's transform class which provides common image transformations. PyTorch also provides functionalities to load and store the data samples with the corresponding labels. In order to create training and validation dataloaders, the in-built DataLoader class was utilized. This wraps an iterable around the dataset, enabling us to easily access and iterate over the data samples in our dataset. The model was configured with the following parameter values-

- Learning rate: 5e-4
- Loss function: CrossEntropyLoss
- Optimizer: Adam optimizer
- Image size: 32
- Patch size: 16 x 16
- Number of classes to classify: 12

Results And Discussion

Table 1 Model results

Model	Validation loss	Validation accuracy
EfficientNet	0.712	0.91
InceptionNet	0.589	0.94
Vision Transformer	0.052	0.984

The above visualizations show the comparison of results of ViT against state-of-the-art CNN architectures on the dataset under study. The ViT based model developed for fruit quality prediction showcased a high accuracy of 98.4. The ViT model was pre-trained on the Image net dataset and fine-tuned. The results below show that ViT performed better than Inception net-based architecture and the EfficentNet-L2 architecture on all the datasets. Both of these models represent the most up-to-date CNN architectures.

ViT requires much less computational resources and training time compared to the other two CNN models.

Conclusions

In this paper, we proposed a deep learning-based machine vision system for grading the fruits based on their outer appearance or freshness. The formulation of an image classification problem as a sequential problem utilizing image patches as tokens and processing it by a transformer is the major engineering component of this study.

Various state-of-the-art deep learning models and methods were applied to the dataset of fruits. Through our experiments, we found that the Vision transformer is the best model for the task. Transfer learning was incorporated with the help of ImageNet pre-trained weights proved to be effective in boosting accuracy. The overall performance of the model to evaluate fruit quality is very good which has met the desired result of the accuracy of about 98.4%. It has been concluded that the Vision transformer has a higher precision rate on a large dataset with reduced training time. Moreover, the application of the vision transformer-based approach has been found to be more accurate in quality assessment as compared to the results previously reported while implementing CNN-based models. Hence, the suggested technique improves the rate of identification of fruit and its freshness detection so that the real-world application demands can be achieved. In the future, the existing database will be enhanced with additional varieties of fruit and vegetable images. Also, the future scope of this implementation will involve the integration of the proposed system that will be helpful in the operation of an automated robotic fruit freshness evaluation system.

Declarations

Conflict of Interest

The authors have no conflicting financial or other interests.

Acknowledgement

The authors wish to acknowledge the research support provided by the Department of Electronics and Communication of College of Engineering Guindy, Anna University.

References

Park & Kim. (2022). *How Do Vision Transformers Work?* ArXiv, abs/2202.06709.

https://doi.org/10.48550/arXiv.2010.11929

Dosovitskiy, Beyer, Kolesnikov, A Weissenborn, Zhai,Unterthiner, Dehghani, Minderer, Heigold, Gelly, Uszkoreit, Houlsby, Neil.(2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,* ArXiv, https://doi.org/10.48550/arxiv.2010.11929

Chen, Xiangning and Hsieh, Cho-Jui and Gong, Boqing. (2021). *When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations*. ArXiv,https://doi.org/10.48550/arxiv.2106.01548

A. Kumar, R. C. Joshi, M. K. Dutta, M. Jonak and R. Burget, Fruit-CNN: An Efficient Deep learning-based Fruit Classification and Quality Assessment for Precision Agriculture, 2021 In: *13th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2021, pp. 60-65, DOI: 10.1109/ICUMT54235.2021.9631643.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision, In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818-2826, DOI: 10.1109/CVPR.2016.308.

S. Fan et al., *On-line detection of defective apples using computer vision system combined with deep learning methods*, J. Food Eng., vol. 286, pp. 110102, Dec. 2020.

S. Chakraborty, F. M. J.M. Shamrat, M. M. Billah, M. A. Jubair, M. Alauddin and R.Ranjan, Implementation of Deep Learning Methods to Identify Rotten Fruits, In: *20215th International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1207-1212, 2021.

John B. Njoroge, Kazunori Ninomiya, Naoshi Kondo and Hideki Toita, *Automated Fruit Grading System using ImageProcessing*, The Society of Instrument and Control Engineers(SICE2002), pp. 1346-1351, August 2002.

P. Deepa, *A Comparative Analysis of Feature Extraction Methods for Fruit Grading Classifications*, International journal of emerging technologies in computational and applied sciences (IJETCAS), vol. 13, no. 138, 2013.

J. Blasco, N. Alexios, E. Molto, *A Machine vision system for automatic quality grading of fruit*, Biosyst. Eng., 85 (4) (2003), pp. 415-423

Mandeep Kaur, Reecha Sharma, *ANN-based Technique for Vegetable Quality Detection*, IOSR Journal of Electronics and Communication Engineering", ISSN: 2278- 8735, pp:62-70, 2015.

Dakshayini Patil, *Fruit disease detection and classification using image processing*. International Journal for Research in Engineering Application & Management (IJREAM). ISSN : 2454-9150, pp:128-131, 2018.

Figures

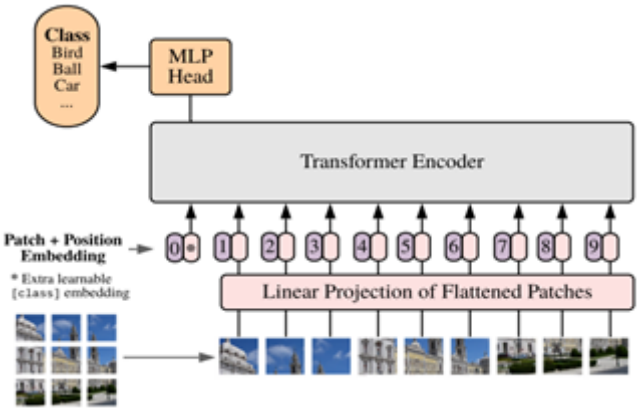

Figure 1

Architecture of Vision Transformer

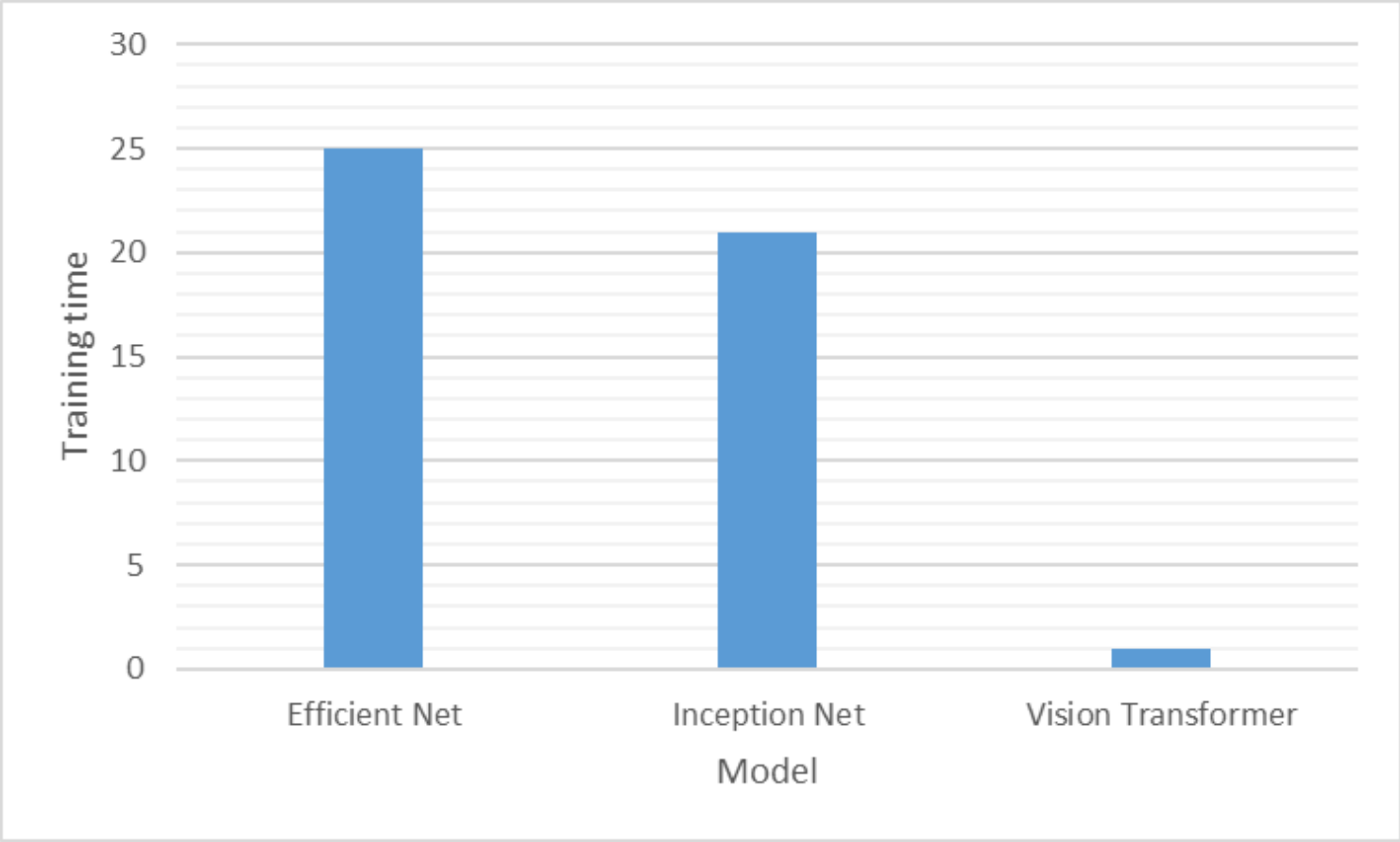

Figure 2

Training duration for different models

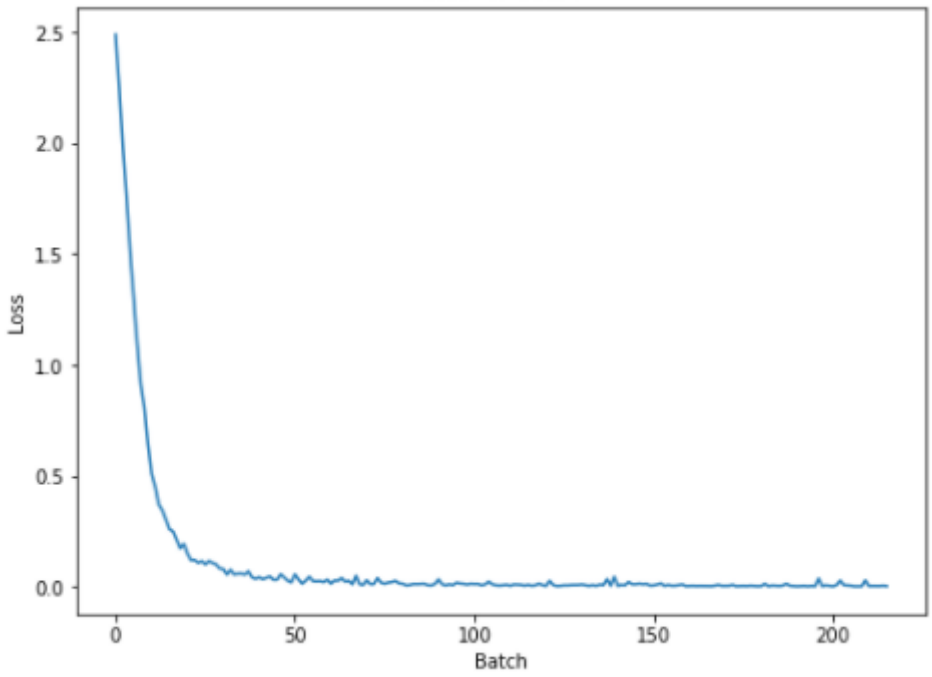

Figure 3

Loss for Batch

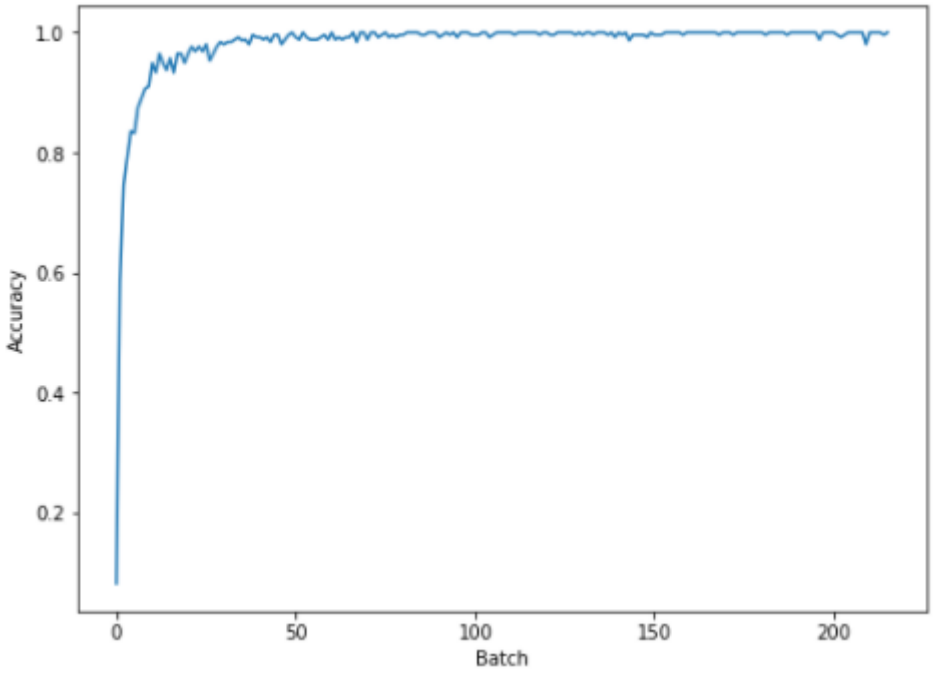

Figure 4

Accuracy for Batch